Instructions

Please upload your project report as a file attachment. The title page should include the title of the project, name of the participants, and a brief abstract. The rest of the report should be typed and double-spaced on standard-sized paper, with 1" margins on all sides. Your report should be approximately 15 pages (with a margin of  pages). If you think it is not possible to fit the material within this page limit, please contact me.

The report should have an introduction with clear description of the problem and (if applicable) references to relevant work on that topic. There should be a methodology section with concise description of the techniques used in the data analysis. This should be followed by the result section where you should discuss the most significant findings (with appropriate tables and plots) of your analysis. If you plan to use multiple algorithms, please create subsections for each of them in the methodology and results sections. The report should conclude with a conclusion section where main takeaways, limitation of the analysis and future direction should be discussed briefly.

Please do not include any R output in the main part of the report. If you want to show results obtained from a R output, please create tables for them to include in the main report. This instruction, however, is not valid for plots. You may put R plots in the main body of the report. Please create a supplementary for the R outputs. The supplementary does not have any page limit. If you want me to look for specific results in the supplementary section, please reference that appropriately in the main report.

A Cluster-Based Approach to Fitting Regression Models

L. Austin Hadamuscin, James Hawkins, Tomiwa Omotesho, Deep Sagar Karki

STAT 6440: Data Mining

Dr. Shuchismita Sarkar

April 23, 2022

Abstract:

Contents

[1 Introduction 4](#_Toc100487228)

[1.1 Motivation 4](#_Toc100487229)

[1.2 Objectives 4](#_Toc100487230)

[2 Data Description 5](#_Toc100487231)

[3 Methodology 6](#_Toc100487232)

[3.1 Leave-One-Out Cross-Validation 6](#_Toc100487233)

[3.2 k-NN 7](#_Toc100487234)

[3.3 k-Means Clustering 7](#_Toc100487235)

[3.4 Model-Based Clustering 8](#_Toc100487236)

[3.5 Random Forest 9](#_Toc100487237)

[3.6 Linear Regression 9](#_Toc100487238)

[4 Results 10](#_Toc100487239)

[4.1 k-NN 10](#_Toc100487240)

[4.2 k-Means 10](#_Toc100487241)

[4.3 Model-Based 10](#_Toc100487242)

[5 Appendix 11](#_Toc100487243)

[6 Citations 12](#_Toc100487244)

# Introduction

## Motivation

## Objectives

# Data Description

# Methodology

## Leave-One-Out Cross-Validation

The relatively small number observations in our data allowed us the luxury of using leave-one-out cross validation (LOOCV) whenever we needed to find on optimal parameter. LOOCV uses the following steps:

* partition all data into a training set except for one observation that will be used as our testing set;
* build a model using the training set;
* use this model to predict the response value in the testing set and calculate the mean square error (MSE);
* and repeat this process until every observation has been used in the testing set.

The test MSE is then calculated by taking the average of the MSE*i* values calculated in the LOOCV process. This process is repeated using different parameters and the set of parameters with the lowest test MSE are our optimal parameters. To expedite this process, we used the caret package (Kuhn, 2021) along with other packages in the statistical programming language R (R Core Team, 2021) (RStudio Team, 2021). Using LOOCV over other methods, such as k-fold cross validation, gives us a low bias estimate of test MSE because our training set contains almost all the data; however, because substantial correlation exists between the MSE*i* from the overlapping in the training and testing sets, the variance of the estimate of test MSE is very high, meaning that any addition of new data could result in vastly different outcomes for our optimal parameters. We chose to accept this variance-bias trade-off because each observation is one country partitioning our data into a train, validation, and test set would mean that whole countries would be excluded from the model building process which greatly reduces the applicability of our final model.

## k-NN

We used k nearest neighbors (k-NN) to impute missing data. k-NN works by calculating the Euclidean distance between each observation in the data set, and each new observation, : where represents the number of predictors. In classification k-NN, the new value takes the class of the majority among its closest observations based on Euclidean distance. In the case of ties, there are several methods used to classify the new observations such as using the majority class among all observations. Our response variable is continuous, so we used regression k-NN where new observations are assigned the average response of the closest observations. We used this concept to impute our missing data. We data mined the optimal by utilizing the train function from the caret R package (Kuhn, 2021) and the knn method from the FNN R package (Beygelzimer, et al., 2019) to perform LOOCV as described in section 3.1. While doing initial research into k-NN data imputation, we found the function knnImputation from the archived R package DMwR (Torgo, 2010). We modified the code of this function to estimate the missing values for each observation, using the mean of the present values from their nearest neighbors.

## k-Means Clustering

We used a -means approach to clustering to cluster our data. The -means clustering algorithm is fairly simple and can be easily explained:

* randomly choose points in -dimentional space to use as the initial centroids;
* calculate the Euclidean distance between the observations and the centroids and assign each observation to its closest centroid;
* recompute new centroids by calculating the center of each group that was created in the previous step. Repeat the previous step;
* The algorithm stops when there is no change to the cluster assignments.

We can measure how well a -means algorithm performs by calculating its within sum of squares () which is a measure of the density of observations in the clusters and is defined as:

where is the set of observations in cluster . The performance of a -means clustering algorithm is highly dependent on the initialization of the centroids created in the first step of the algorithm so to combat this, we randomly generated a set 1000 seeds from a uniform(0,900000) distribution, using the ceiling function (R Core Team, 2021) to round to the next highest whole number. We used theses seeds to run 1000 iterations of the algorithm, choosing the seed that produced the lowest . Using the same 1000 seeds, this process was repeated for and the model that resulted in the lowest without the increase in k resulting in a negligibly lower was chosen.

## Model-Based Clustering

We also took a model-based approach to clustering our data. Model-based clustering uses finite mixture models which are a linear combination of weighted density functions:

where are the weights, , , and is the mixture order, or number of components, that use the Expected-Maximization (EM) iterative algorithm to estimate . Like the k-means clustering algorithm, model-based clustering algorithms are highly dependent on initialization so many models must be run to get the optimal assignment of clusters. To determine what set of clusters are the best, a modified version of the likelihood, Bayesian Information Criterion (BIC), is used:

where is the maximum likelihood and is the number of parameters to be estimated. We used the Mclust function from the mclust R package (Scrucca, Fop, Murphy, & Raftery, 2016) to perform the clustering. This package automatically runs the algorithm at different initializations and chooses the best model and returns its negative BIC. We ran the function using K and chose the model with the maximum negative BIC without the increase in K resulting in a negligibly higher negative BIC. We then compared this model, visually, with the model we created using k-means and chose the one that we felt created the best clusters.

## Random Forest

The necessity of ensemble techniques comes with the demand of reduction of noise in the

predictive model. The ensemble technique uses multiple models and takes the average of them.

This procedure helps to reduce the variances while controlling the bias . One of the

preferred ensemble techniques is random forest.

Random forest was first proposed by Leo Breiman from the university of California in 2001.

The random forest consists of multiple decision tress which are formed in such a way that each

tree has no correlation with every other trees. These independent trees are used either for

regression or classification problem.

The following are the steps used in Random Forest:

(i) Fix a proper value of predictors usually labelled as “m”.

(ii) Select a new subset of predictors from the whole set of predictors depending on “m”

(iii) Train the data set with the feature subset to create decision tree for each group of training set

(iv) Choose a new and repeat the process above until travel all the feature subsets.

(v) Test the prediction on the test data set.

Random forest consists of large number of decision trees. The randomness is incorporated in the

model building process, including the selection of samples subset and feature subset which

guarantees the independence of each decision tree. With the method of bagging where random

boot strap samples are selected ensures the randomness of samples selection. Similarly, with the

help of bagging subsets of features are selected based on the value of “m” from the entire

feature set. The influence of “m’ is highly noted on the performance of random forest. With the

increase in the value of “m”, the correlation between each tree in the training model can be

improved to make the trees independent.

For our project, we fitted the random forest on each of the clusters and extracted the “variable importance” from four different models. Since, we had 11 predictors in all of our clustered data we set the range of “m” predictors to be for each clusters. The presence of small number of observations led us to use Leave One Out Cross-Validation folds for each clusters for deciding the optimal number of “m” predictors.

The number of optimal “m” was selected based on the corresponding RMSE. The table below shows the optimal number of “m” and corresponding RMSE on each cluster when LOOCV was implemented:

|  |  |  |
| --- | --- | --- |
| Cluster Number | Optimal “ m” tries | RMSE |
| 1 | 10 | 1.017949 |
| 2 | 6 | 2.400911 |
| 3 | 10 | 2.003527 |
| 4 | 8 | 2.451414 |

In R, “caret” package , for method = “rf” in “train(response~, data,method= “rf”, tune.grid=train\_control,tuneGrid=tune.grid)” was used to build the decision tree models on each of cross validation folds that resulted on an average model.

## Linear Regression

We wanted to see if there is any linear association between the response variable and

the predictors. For this purpose, we considered multiple linear regression model.

Th model assumption for multiple linear regression are:

1. Errors are normally distributed
2. Constant variances
3. The relationship between predictors and response should be linear
4. The observations should be independent of each other

We often encounter the problem of model misspecification with the multiple linear regression.

The model is prone to using extraneous variables which will create the problem of overfitting.

It is also the case that the model lacks some important variables. In order to avoid

these situations and obtain correctly specified models, we look for variable selection

techniques. One technique that we used for variable selection is regularized regression

technique. In a regularized regression technique, the regression coefficients are estimated by

minimizing SSE+P, where SSE represents the sum of square deviations and P is the penalty

factor. In regularized regression the estimated effects are slowly shrunken towards zero. The

regularized regression has two types:

(i) Ridge regression : minimizes SSE+

(ii) Lasso regression. Minimizes SSE+

is called the tuning parameter, and as the tuning parameter becomes very large, the penalty

factor becomes large that forces the regression coefficients to become zero. This process results

in reducing the variance, but bias is compromised unlike in ordinary least square regression.

In R, “glmnet” package is used for regularized regression. Since, regularized regression,

requires the features to be standardized, “glmnet” package does the standardization for us. The

“alpha” parameter in the package was specified to use lasso regression by setting alpha=1.

The is chosen based on the RMSE that corresponds to the minimum.

Once, the variables are selected the model is fitted with the appropriate regressors. It is also

important that our model does not suffer from potential multicollinearity. So, once the model is

fitted we check the multicollinearity in the model using VIF (variance inflation factor) from the

“car” package. The VIF values higher than 4 signifies that the model suffers from moderate

multicollinearity, whereas the VIF higher than 8 signifies that the model suffers from severe

multicollinearity. If the problem of multicollinearity exists then we will drop the highly

correlated variables to get rid of the multicollinearity from the existing model. The model

without multicollinearity is used for the prediction purposes. The significance of the regressors

in the model will be compared with the significance level of =0.05.

# Results

## k-NN

## k-Means

Following our methods in section 3.3, we ran the k-means algorithm at various values of using the same set of 1000 seeds for each and displayed the results in Figure 4‑1.

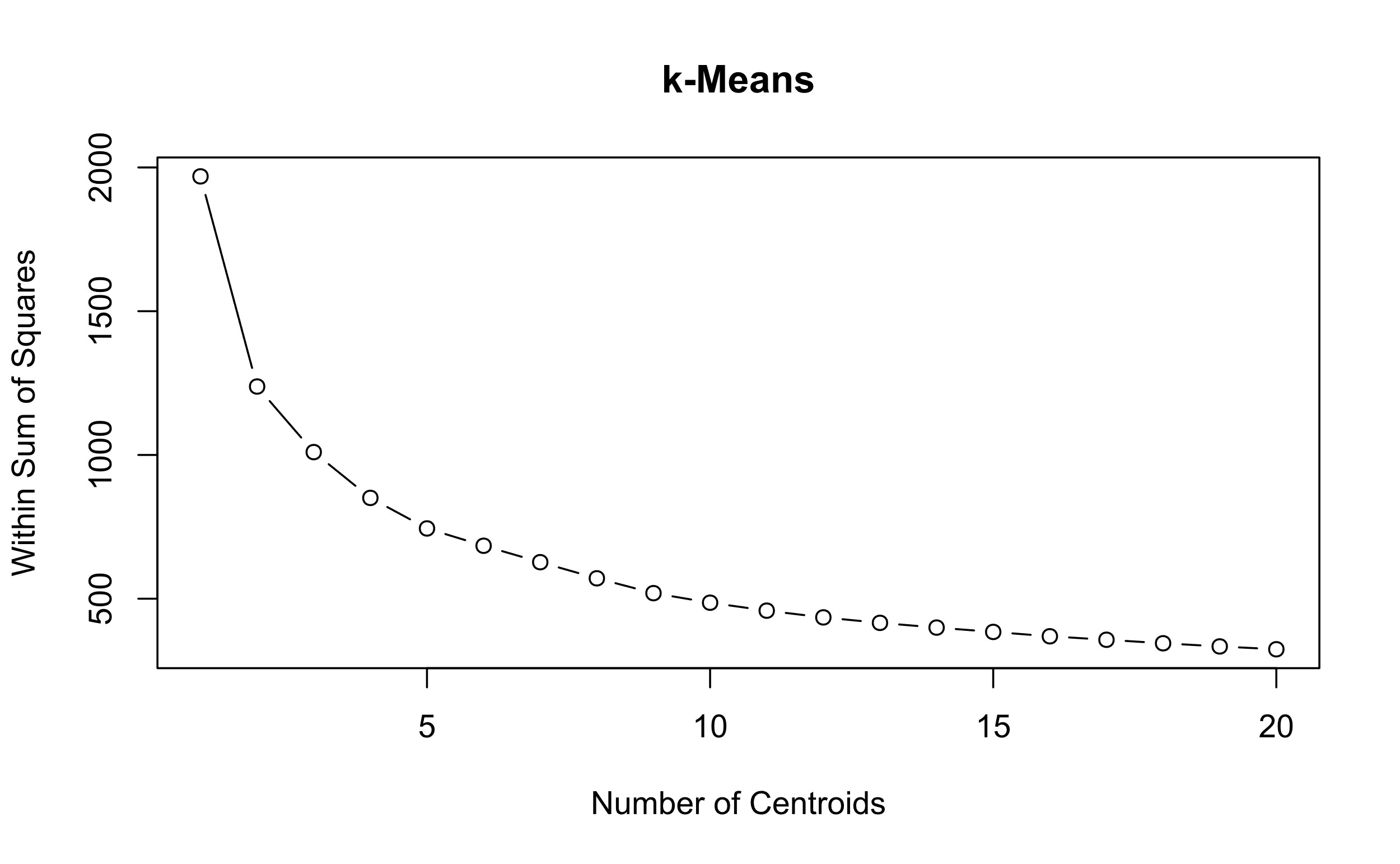


Figure 4‑1 The within sum of squares results for the k-means clustering algorithm at different k

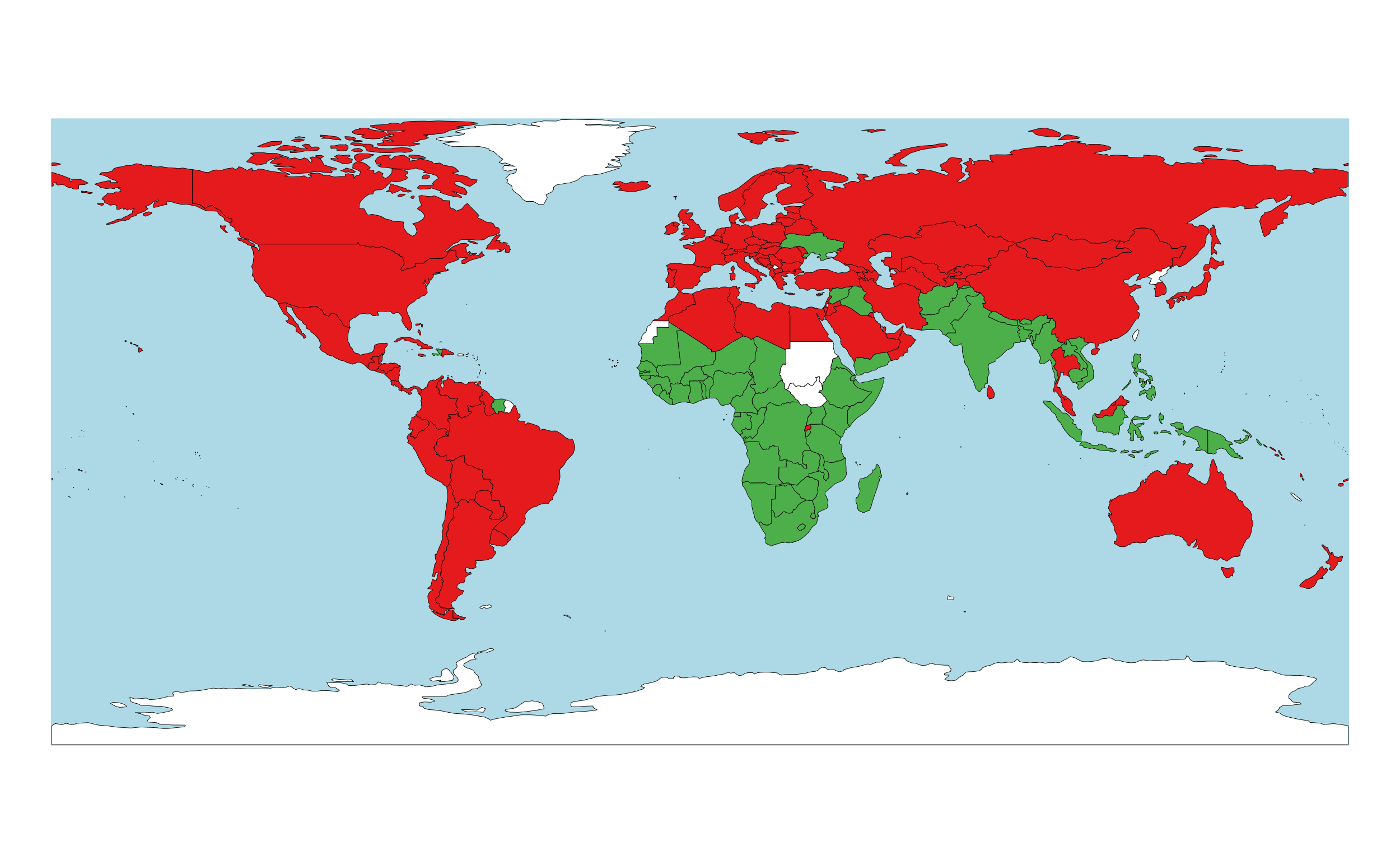
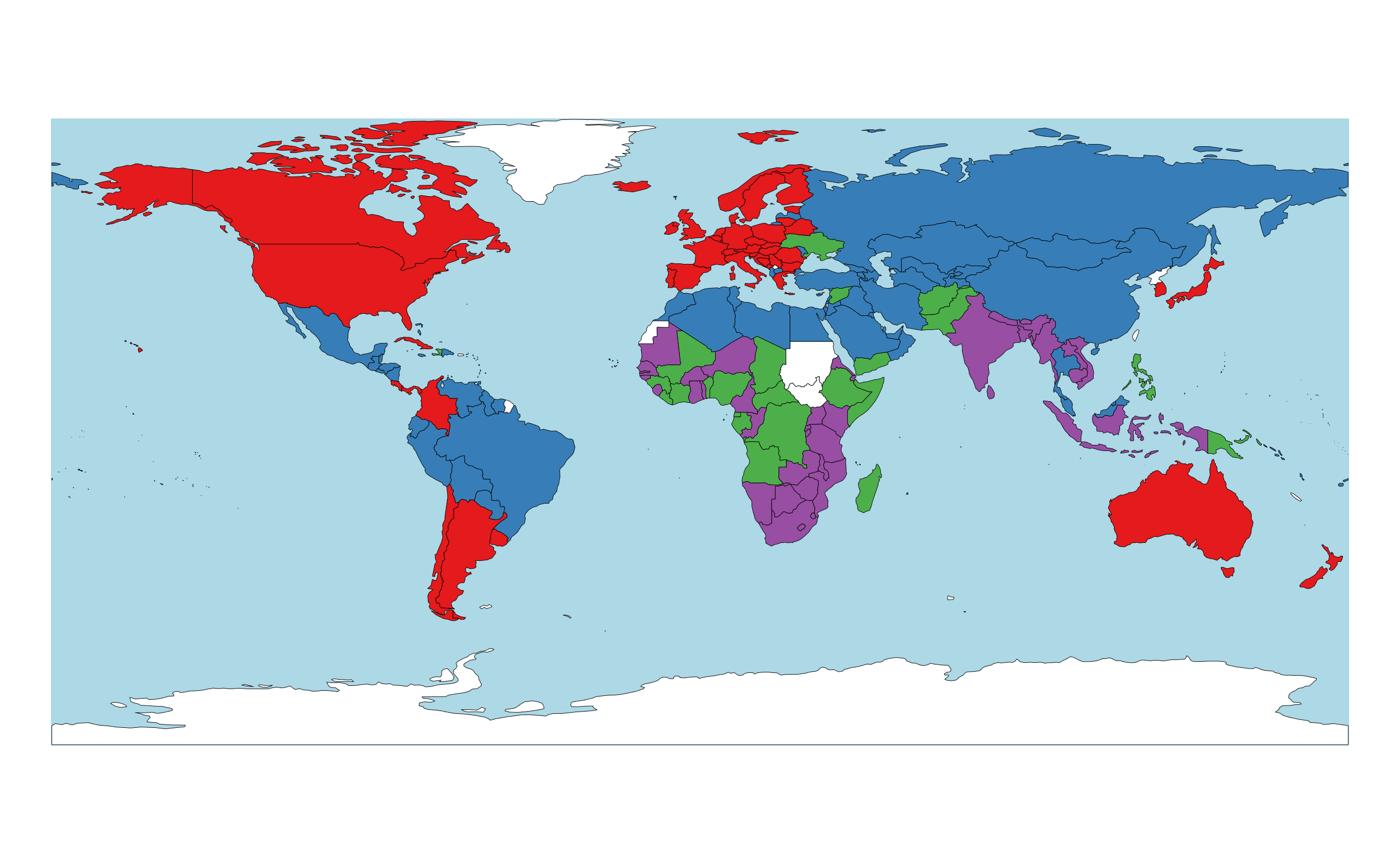
We can see that there is not a distinct “elbow” in the graph that can be chosen as our optimal other than a slight bend at = 2 which results in the map seen in figure 4‑2. The algorithm did a decent job at clustering the countries; however, there issues that come with grouping the countries in such a broad manner. For example, less economically developed countries such as Turkey and Ecuador are grouped with more economically developed countries such as Germany and South Korea. We looked at = 4 to help break up some of these oddities and because the reduction in the within sum of squares beyond this point was negligible. The resulting map can be seen in Figure 4‑3

Figure 4‑2 Visualization of the clusters when k-means is used and k = 2



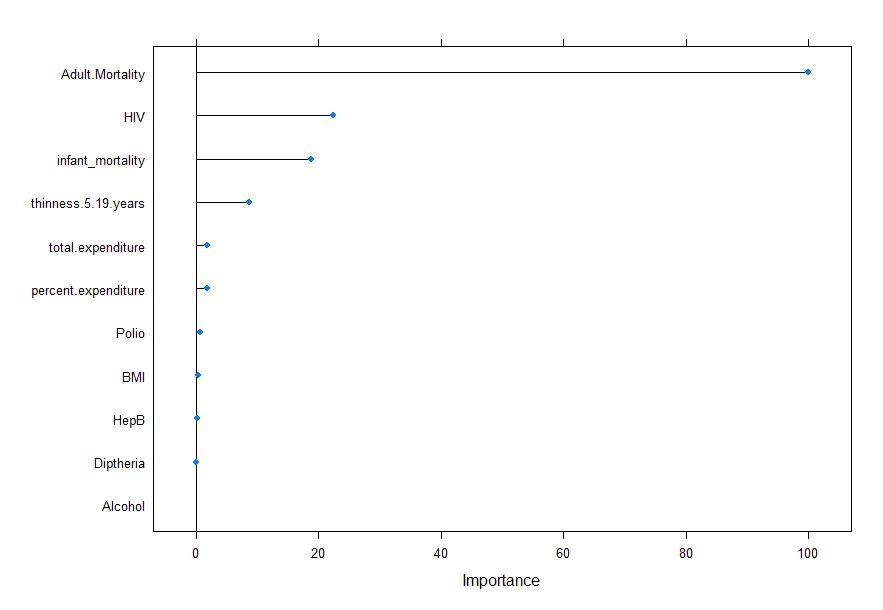
## Model-Based

## K-Means vs Model-Based

## Random Forest

### Cluster 1

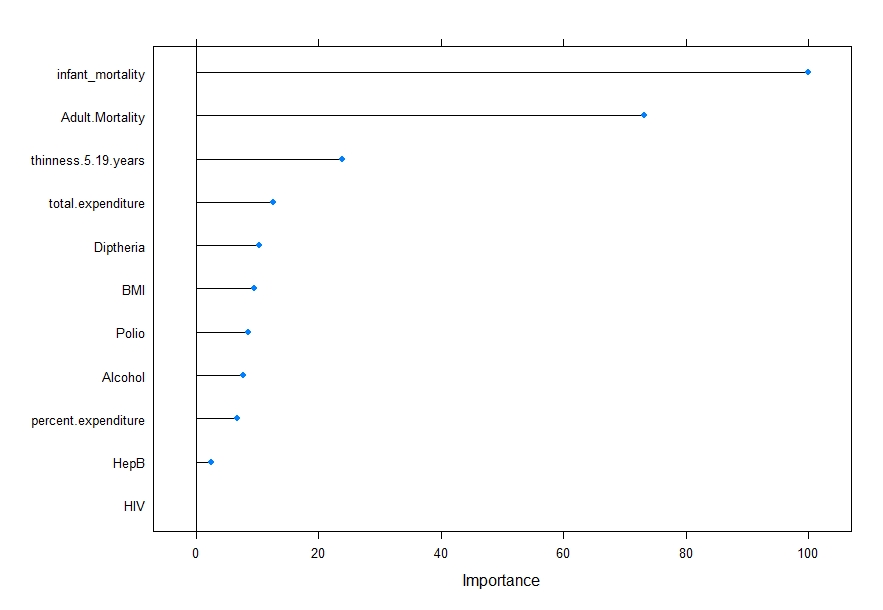
After random forest was fitted into cluster 1, it was found that “Adult mortality” was ranked as a powerful predictor outranking all other variables followed by “HIV”, and “Infant Mortality” etc. The complete ranking of the variable importance is given below:



The variables ranked tells us that the countries in Cluster 1 had high problems with the adult mortality and the prevalence of HIV. Based on the data of 2014, it would be recommended to consider the first four factors and be highly prioritized by the government to improve the life expectancy of the population.

4.5.2 Cluster 2

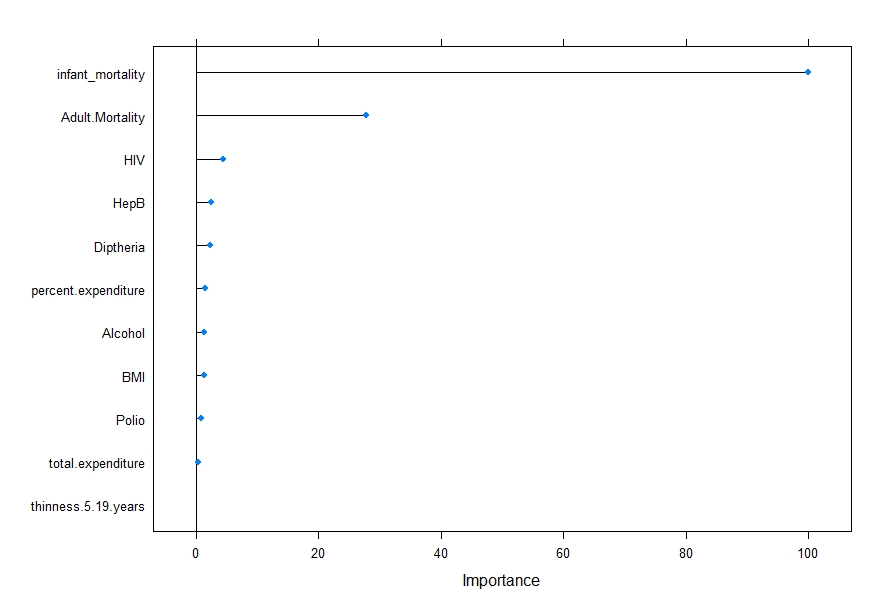
The similar approach was used for cluster 2. Upon inspection of the variable importance, it was found that “infant mortality” was ranked as a powerful predictor outranking all other variables closely followed by “adult mortality”. The complete ranking of the variable importance is given below:



The variables ranked tells us that the countries in Cluster 2 have high mortality rate among the population. The thinness among children for age group between 5 to 19 years is also a serious issue. The government and other institutions should increase percent of the expenditure spent on the population to alleviate the problems . The situation in terms of population mortality is worse in cluster 2 compared to cluster 1 according to 2014 data.

4.5.3 Cluster 3

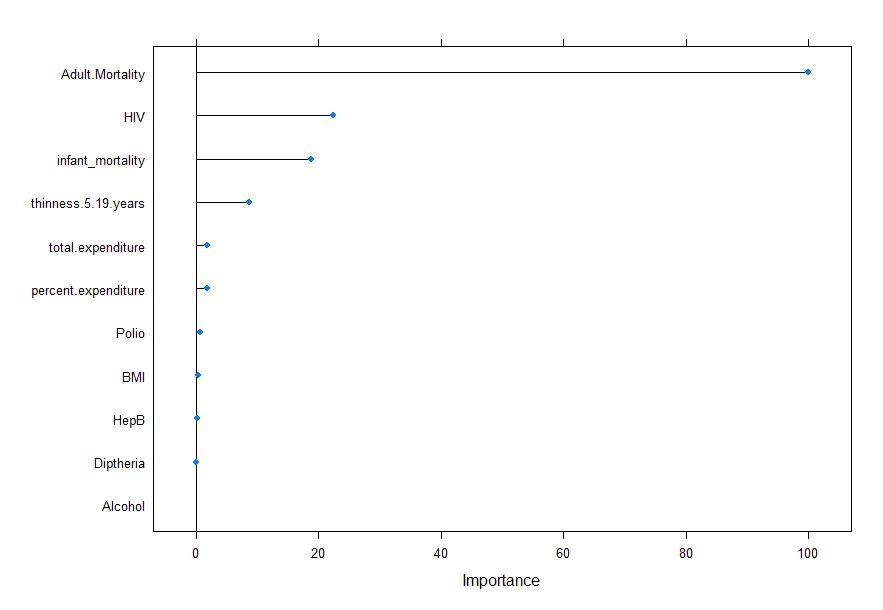
For cluster 3 , “infant mortality” was ranked as a powerful predictor outranking all other variables closely followed by “adult mortality”. “HIV” . The complete ranking of the variable importance is given below



. The ranking of first two variables is similar in cluster 2. However, thinness among children aged 5 to 19 years seemed to have no contribution in the prediction unlike in cluster 2. Countries in cluster 3 should also heavily focus on decreasing the infant and adult mortality. It is also recommended to consider the control such as decreasing the prevalence of HIV, Hepatitis B and Diphtheria to improve the life expectancy.

4.5.4 Cluster 4

For cluster 4, “Adult mortality” was ranked as a powerful predictor outranking all other variables followed by “HIV”, “infant mortality, and “thinness.5.19. years” etc. The complete ranking of the variable importance is given below



Countries in cluster 4 share the same first three predictors as in cluster 1 and 3. Countries in this cluster should follow the similar approach as mentioned earlier for cluster 1 and cluster 3.

4.6 Linear Regression

4.6.1 Cluster 1

In cluster 1, the model after lasso regression comprises of the following predictors:

1. BMI (ii) Alcohol (iii) Adult Mortality (iv) HepB (v) precent. expenditure (vi) Polio

(vii) thinness.5.19.years (viii) total expenditure. The coefficients corresponding to the three

predictors were shrunk to 0 and hence were not included in the final model. The variance

inflation factor (VIF) was used from the “car” package to check the multicollinearity among the

regressors. We noticed that the model that included the 8 regressors above did not suffer from

multicollinearity.

The constant variance assumption was satisfactory. Similarly, normality of

residuals were also satisfactory except for wo regressors HepB and Polio. We tried different

transformations but there was not much improvement in the normality. Overall, the final model

consisted of the 8 regressors above with the adjusted R-squared value of 0.9039. BMI, Adult

mortality, percent expenditure, Polio thinness.5.19 years and total expenditure were all found to

be significant in the model The results of assumptions checking and summary of the regression

model are attached on the appendix>

4.6.2 Cluster 2

In cluster 2, the model after lasso regression comprises of the following predictors:

1. BMI (ii) Alcohol (iii) Adult Mortality (iv) HepB (v) precent. expenditure (vi) Polio

(vii) thinness.5.19.years (viii) total expenditure (ix) Diphtheria (x) HIV (xi) infant mortality .

The variance inflation factor (VIF) was used from the “car” package to check the

multicollinearity among the regressors. We noticed that the model that included the regressors

above suffered from multicollinearity . Diphtheria and HepB had high VIF values of 55.68 and

48.831 respectively. So, we dropped these two variables and fit the new regression model. The

VIF for the new regression model then did not suffer from the multicollinearity.

The constant variance assumption was satisfactory. Similarly, normality of residuals were good

except for two regressors HIV and Polio. We tried different transformations but there was not

much improvement in the normality. Overall, the final model consisted of the 9 regressors

above with the adjusted R-squared value of 0.7269. Alcohol, Adult Mortality, and infant

mortality were found to be significant regressors in the model. The results of assumptions

checking and summary of the regression model are attached on the appendix>

4.6.3 Cluster 3

In cluster 3, we used the lasso regression and the model comprises of the predictors:

(i) Alcohol (ii) Adult. Mortality (iii) HIV (iv) infant mortality (v) precent expenditure (vi)

Polio (vii) total expenditure. The variance inflation factor (VIF) was used from the “car”

package to check the multicollinearity among the regressors. The variables HIV, precent.

expenditure and total expenditure with the corresponding VIF values of 5.375702, 5.812510

and 6.535088 showed the presence of moderate multicollinearity in the model. So, we dropped

the two variables total expenditure and HIV and fitted the model again. The final model now

did not suffer from multicollinearity. The constant variance assumption was satisfactory.

Similarly, normality of residuals was satisfactory. The final model consisted of 5 regressors

with the adjusted R-squared value of 0.8498. Infant mortality was the only significant regressor

in the model. The results of assumptions checking and summary of the regression model are

attached on the appendix>

4.6.4 Cluster 4

In cluster 4, we used the lasso regression and the final model comprises of the predictors:

(i) BMI (ii) Adult Mortality (iii) HepB (iv) HIV(v) Infant mortality (vi) percent expenditure

(vii) total expenditure. The VIF values of 6.902179 and 4.911363 for Adult Mortality and HIV

showed the presence of multicollinearity. So, we dropped the variable HIV and fit the model

again. Upon checking the VIF, we noticed that model did not suffer from multicollinearity.

The constant variance assumption was satisfactory. Similarly, Normality of residuals were also

satisfactory. The final model consisted of the 6 regressors with the adjusted R-squared value of

0.8939. Adult mortality, percent expenditure and total expenditure were found to be significant

regressors in the model. The results of assumptions checking and summary of the regression

model are attached on the appendix>

# 6Appendix

# 7Citations

Parmar, A., Katariya, R., & Patel, V. (2018, August). A review on random forest: An ensemble classifier. In *International Conference on Intelligent Data Communication Technologies and Internet of Things* (pp. 758-763). Springer, Cham.

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., & Li, S. (2019). FNN: Fast Nearest Neighbor Search Algorithms and Applications. *R package version 1.1.3*. Retrieved from https://CRAN.R-project.org/package=FNN

Kuhn, M. (2021). caret: Classification and Regression Training. R package version 6.0-90. Retrieved from https://CRAN.R-project.org/package=caret

R Core Team. (2021). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

RStudio Team. (2021). RStudio: Integrated Development Environment for R. Boston, MA. Retrieved from http://www.rstudio.com/

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal, 8*(1), 289-317. doi:https://doi.org/10.32614/RJ-2016-021

Torgo, L. (2010). Data Mining with R, learning with case studies. Chapman and Hall/CRC. Retrieved from http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR